# MVReward: Better Aligning and Evaluating Multi-View Diffusion Models with Human Preferences

**Weitao Wang**[1*], **Haoran Xu**[2*], **Yuxiao Yang**[1], **Zhifang Liu**[1], **Jun Meng**[2†], **Haoqian Wang**[1†]

[1]Tsinghua Shenzhen International Graduate School, Tsinghua University
[2]Zhejiang University

## Abstract

Recent years have witnessed remarkable progress in 3D content generation. However, corresponding evaluation methods struggle to keep pace. Automatic approaches have proven challenging to align with human preferences, and the mixed comparison of text- and image-driven methods often leads to unfair evaluations. In this paper, we present a comprehensive framework to better align and evaluate multi-view diffusion models with human preferences. To begin with, we first collect and filter a standardized image prompt set from DALL·E and Objaverse, which we then use to generate multi-view assets with several multi-view diffusion models. Through a systematic ranking pipeline on these assets, we obtain a human annotation dataset with 16k expert pairwise comparisons and train a reward model, coined MVReward, to effectively encode human preferences. With MVReward, image-driven 3D methods can be evaluated against each other in a more fair and transparent manner. Building on this, we further propose Multi-View Preference Learning (MVP), a plug-and-play multi-view diffusion tuning strategy. Extensive experiments demonstrate that MVReward can serve as a reliable metric and MVP consistently enhances the alignment of multi-view diffusion models with human preferences.

**Code** — https://github.com/victor-thu/MVReward

## 1 Introduction

3D content generation is developing rapidly, thanks to the powerful generation ability of 2D diffusion and increasing efforts (Poole et al. 2022; Lin et al. 2023) to lift its rich priors to 3D. Given a text or image prompt, current methods can generate 3D objects with detailed geometry and texture in seconds, with multi-view diffusion models playing an indispensable role. These models pave the way for fast and robust 3D generation by training view-aware diffusions to generate consistent images. As the 3D generation field gains more attention, its further development is highly anticipated.

Nonetheless, the lack of corresponding 3D evaluation methods is becoming a significant obstacle to progress in this area. As fully illustrated in previous research (Otani et al. 2023; Wu et al. 2024), existing automatic methods such as FID, LPIPS, CLIPScore, etc. often fail to align with human preferences, as shown in Figure 1. Furthermore, taking the widely-used GSO dataset (Downs et al. 2022) as an example, the test objects are not consistently unified across different methods, exacerbating this misalignment. Meanwhile, comparing generated content to ground truth (GT) data in an ill-posed task, is a one-sided approach that may stifle the model's creative potential. The everyday objects in the GSO dataset also lack adequate evaluation of those imaginative, unrealistic objects that are common in generative tasks. Considering the reasons above, most 3D generation methods rely heavily on qualitative analysis—comparing results through the subjective judgment of the viewer which may differ greatly across individuals.

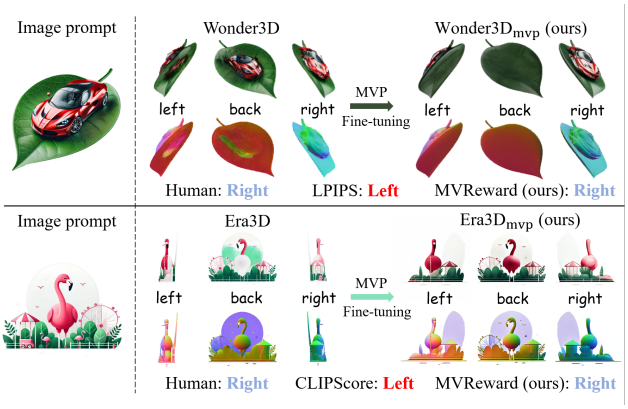


Figure 1: Automatic metrics often struggle to align with human preferences in evaluating image-to-3D tasks. Our MVReward model fills this gap and our MVP further enhances the alignment of existing multi-view diffusion models with human preferences.

Several works (Xu et al. 2024b; Wu et al. 2023) attempt to apply RLHF (Reinforcement Learning from Human Feedback) methods to text-to-3D tasks to solve the problem of misalignment with human preferences. However, similar research is absent in the image-to-3D field, particularly for multi-view diffusion methods, and the evaluation biases brought by the GSO dataset have not been taken seriously.

On the other side, the mixed evaluation of text- and

*These authors contributed equally.
†Corresponding Authors.

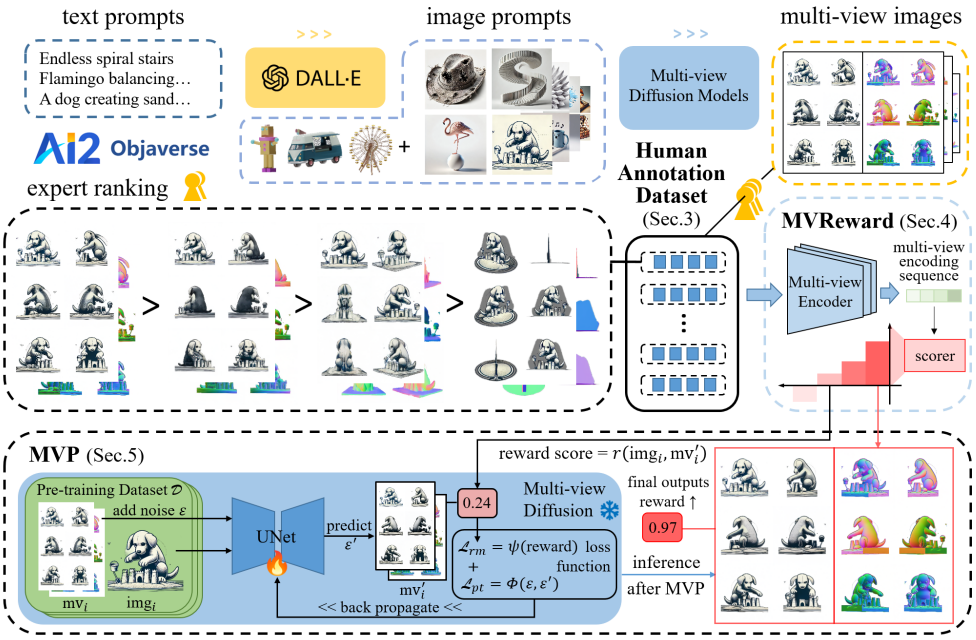


Figure 2: An overview of our whole framework. Our human annotation dataset is constructed by a text prompt ⇒ image prompt ⇒ multi-view images ⇒ human annotation procedure (Sec.3). Then we train our MVReward model, which includes a multi-view encoder and a scorer to effectively encode human preferences and evaluate multi-view images (Sec.4). Finally we propose MVP to fine-tune multi-view diffusion models by combining pre-trained loss with our reward loss (Sec.5).

image-driven methods is also becoming a serious issue. Image-to-3D tasks have distinct application areas from text-to-3D, and images, as more fine-grained inputs, require higher quality than text. Some evaluation methods (Wu et al. 2024) simply use the text-to-image results of text prompts as inputs for image-driven methods like (Long et al. 2024) during 3D evaluation. Since image-driven methods typically remove the inputs' background for better generation results, we observe incomplete objects and missing scenes like the example in Figure 3, causing potential unfair comparisons between text- and image-driven methods.

In light of these challenges, in this paper we introduce a complete framework specially designed for multi-view diffusion based methods, providing a normalized image-to-3D evaluation environment and better alignment with human preferences. We begin by generating and filtering a standardized image prompt set from DALL·E (Ramesh et al. 2021) and Objaverse (Deitke et al. 2023), ensuring the object(s) in each image are fully visible with well-designed geometry and texture. Then we select four multi-view diffusion methods to generate RGB and normal multi-view assets using image prompts. After annotating and collecting 16k expert pairwise comparisons of these multi-view assets as a human annotation dataset, we train the MVReward model on it. We further propose MVP, a plug-and-play tuning strategy to align multi-view diffusion models with human preferences.

Our main contributions are as follows:

- We systematically analyze the chaos and challenges in the evaluation of image-driven 3D methods, and based on this, create a comprehensive pipeline that includes filtering a standardized image prompt set, generating high-quality multi-view assets, and collecting 16k expert pairwise comparisons as a human annotation dataset.
- We train MVReward, the first general-purpose human preference reward model for multi-view diffusion, which can serve as a reliable image-to-3D metric.
- We present MVP, a plug-and-play multi-view diffusion tuning strategy, consistently enhancing the alignment of multi-view diffusion models with human preferences.

## 2 Related Work

### 2.1 3D Generation with Multi-view Diffusion

Current 3D generation techniques can be broadly classified into three categories: distillation based methods, multi-view based methods and feed-forward methods. Distillation based methods (Wang et al. 2023, 2024a; Chen et al. 2023) aim to leverage the strong priors of 2D diffusion models to generate consistent 3D representations like NeRF by designing a Score Distillation Sampling (SDS) loss, but suffer from low efficiency and multi-face problem. Multi-view based methods (Liu et al. 2023a,b) alleviate this issue by directly fine-tuning 2D diffusion models to generate multi-view images without relying on 3D representation and time-consuming per-shape optimization. Feed-forward methods (Li et al. 2023; Hong et al. 2023; Tang et al. 2024; Wang et al. 2024b; Xu et al. 2024c,a) seek to generate 3D shapes end-to-end in seconds, by integrating off-the-shelf multi-view diffusions with sparse-view large reconstruction models. It is evident

that nowadays multi-view diffusion plays a significant role in 3D generation. Zero123 (Liu et al. 2023a) first trains a diffusion model conditioned on the reference image and camera viewpoint, pioneering the multi-view generation. More advancements (Chen et al. 2024; Voleti et al. 2024) are emerging after it and the potential of the multi-view diffusion in 3D generation continues to be explored.

### 2.2 Learning from Human Feedback

Pre-trained generative models often exhibit misalignment with human intent due to the limited size and inherent biases of the training data. Reinforcement Learning from Human Feedback (RLHF) is a commonly used method to address this deviation. Natural language processing (NLP) methods (Christiano et al. 2017; Ibarz et al. 2018) first align the language model with human preference via training a reward model (RM). InstructGPT (Ouyang et al. 2022) further applies RLHF to GPT-3, improving its performance significantly in multi-task NLP. ImageReward (Xu et al. 2024b) and Pick-a-pic (Kirstain et al. 2023) expand the application scope of RLHF to text-to-image generation by collecting human annotation image datasets and training RMs on them. RLHF methods in 3D generation (Ye et al. 2024) are also emerging, but research on aligning multi-view images with human preference remains blank and our MVReward fills it.

### 2.3 3D Generation Evaluation

Evaluating 3D generation has long been a challenging task, requiring rich 3D priors and a deep understanding of physical properties. Existing 3D methods typically rely on automatic metrics along with user studies for comparison. The GSO (Downs et al. 2022) and Omni3D (Brazil et al. 2023) datasets are widely used to measure the distance between the generated novel views (or 3D shapes) and the reference ones using PSNR, SSIM, LPIPS (or Chamfer Distance, Volume IOU). In the absence of a reference set, multimodal pre-trained embeddings such as CLIP (Radford et al. 2021) and BLIP (Li et al. 2022) may also be utilized to calculate similarities between different views. Recent work (Wu et al. 2024; He et al. 2023) develop several GPT4V (Achiam et al. 2023) intervention procedures to leverage its strong priors in 3D perception. Concurrently, RLHF methods (Ye et al. 2024; Choi et al. 2024) seek to train reward models for text-to-3D evaluation. Different from the above, our method proposes a standardized evaluation process for image-driven 3D methods, especially the multi-view based ones.

## 3 Human Annotation Dataset

Our whole framework is clearly overviewed in Figure 2. In this section, we elaborate on the construction of our human annotation dataset, which is prepared for our MVReward training and future research in human preference learning.

### 3.1 Text Prompts

Text prompts remain essential in image-to-3D tasks by controlling the details of image prompts using text-to-image models. In practice, without advanced knowledge of the potential image input, we often simulate the image prompt distribution by crafting suitable text prompts.

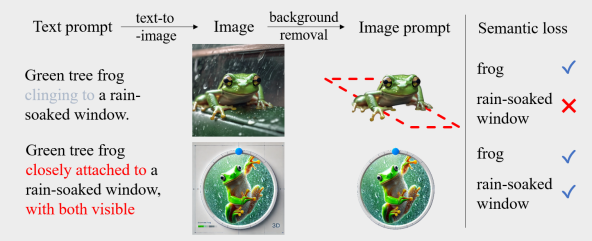


Figure 3: Example of our text prompt enhancements to prevent potential semantic loss brought by the background removal of image-to-3D methods.

Our text prompt set is mainly derived from text-to-3D evaluation methods (He et al. 2023; Wu et al. 2024), with necessary enhancements due to the inherent biases in text- and image-to-3D tasks. For instance, the background of the input image is typically removed by image-to-3D methods to ensure the object-centric generation quality. Therefore, for text prompts with background descriptions, we delete or edit them by adding strong conjunctions like *'closely attached to ... with both visible'* instead of *'clinging to'* to prevent semantic loss of the image prompts during background removal (see Figure 3). In addition, to emphasize the geometry and texture that are of interest in image-to-3D tasks, text prompts are filtered and more generated with increasing complexity and creativity in mind, leveraging GPT4V (details in Appendix B.1).

After a clustering algorithm (Van der Maaten and Hinton 2008) standardizing and unifying the distribution of text prompts, we finally obtain 600 high-quality text prompts for image prompt generation.

### 3.2 Image Prompts

Image prompts are crucial for both generation and evaluation in the image-to-3D task. As discussed in Section 1, the absence of a standardized image prompt set hampers fair evaluation, further limiting 3D generation progress. To solve this issue, we generate image prompts for two key purposes: constructing a human annotation dataset for training an efficient and robust reward model, and providing a standardized set for fair comparisons across image-to-3D methods.

Our image prompt set is expected to exhibit increasing complexity and creativity in geometry and texture, owing to our complicated construction of the text prompt set. For example, *a gift wrapped with mysterious symbols* is a common object with simple geometry but complex texture, whereas *a swan with feathers resembling origami folds* is a creative object with intricate geometry but plain texture. We use the high-quality text-to-image model DALL·E (Ramesh et al. 2021) to generate samples and carefully select those suitable for multi-view tasks (see Appendix B.2). Figure 4 demonstrates some examples from our image prompt distribution. For each text prompt, we select two distinct samples: one for the reward model training and one for the standardized set.

Meanwhile, we include the same number of non-generative, real-world objects from Objaverse (Deitke et al. 2023) to increase the diversity of the dataset. We filter the


Figure 4: Examples from the image prompt distribution. Images closer to the edges represent objects with more complex and creative geometry or texture, while those near the center are mainly simple and common.

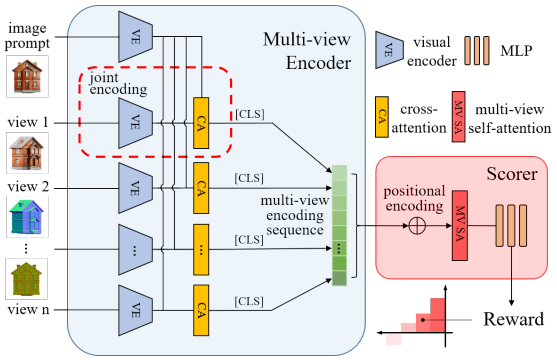


Figure 5: MVReward architecture with a multi-view encoder and a scorer to encode and predict human preferences.

front render view of each Objaverse object as its image prompt. In the end, we get an image dataset with 1200 images as input to multi-view diffusion models and a standardized image prompt set with 600 images for fair comparison.

### 3.3 Human Annotation Pipeline

To grant our reward model with sufficient understanding of multi-view images quality, we construct a human-annotated multi-view images dataset. Firstly, we choose four high-quality multi-view diffusion methods: Wonder3D (Long et al. 2024), Zero123++ (Shi et al. 2023), Envision3D (Pang et al. 2024) and Era3D (Li et al. 2024) following the rule that each method can generate at least six views of RGB and normal images from an image prompt. Although camera systems and poses vary slightly across methods, our hypothesis, supported by empirical results, is that the reward model can evaluate multi-view images under different camera settings just like humans. For an image prompt, we generate two multi-view assets per method using different inference steps (20, 40), resulting in 10200 multi-view assets (122k images) in total, including rendered views from Objaverse.

The next step is to annotate these assets to capture human preferences. We develop a detailed annotator guideline in advance, including task objectives, evaluation criteria, etc. (see Appendix C). 20 annotators are carefully gathered, all with at least a bachelor's degree and an interest in 3D generation. The annotators rank 4-5 multi-view assets from the same image prompt, allowing ties if quality is close. Each annotator is allowed to rank up to 400 asset-lists to minimize fatigue and personal bias. We also provide our own annotations as *researcher annotations* for double checking.

The above annotation pipeline yields 7692 valid rankings. For conflicting annotations, we use the borda count to determine final rankings. At last, we obtain the rankings for 1200 multi-view asset-lists, resulting in 16k valid comparison pairs for the reward model training (except for ties).

## 4 MVReward: Encoding Human Preferences

We decompose the problem of evaluating multi-view images with human preferences into three sub-problems:

- Evaluating the quality of the generated images themselves.
- Assessing the degree of alignment between the generated images and the input image prompt.
- Calculating the spatial and semantic consistency across the generated images.

To this end, we propose MVReward, which incorporates two core modules: a multi-view encoder for sub-problems 1 and 2 by encoding the full representation of a 3D object, and a scorer for sub-problem 3 by capturing and evaluating the inner connections between different views.

### 4.1 Architecture Design

Figure 5 illustrates the architecture of MVReward. We adapt BLIP (Li et al. 2022) as the backbone for our multi-view encoder. The vanilla BLIP model encodes both image and text inputs into a unified representation space, capturing similarities between the two modalities. We extend this approach to connect the input image prompt with the generated views, treating them as distinct modalities.

The multi-view encoder module comprises multiple visual encoders, each corresponding to a different view, but sharing the same parameters. These encoders independently extract features from the image prompt and the generated views. We then apply a cross-attention layer to integrate these features, resulting in a joint encoding for each input-generated view pair.

For each view pair encoding, a special [CLS] token is placed at the beginning of the joint encoding to represent the global feature. These tokens are extracted to form a multi-view encoding sequence, which serves as a comprehensive representation of the views from a 3D asset (e.g., a set of 6×[RGB + normal] images generated by a multi-view diffusion model paired with their corresponding input views worth a sequence with 12 tokens).

The scorer module processes this multi-view encoding sequence, using positional encoding to distinguish between different views and domains (such as RGB and normal). The sequence then undergoes a multi-view self-attention layer to capture correlations between views. Each token from the processed sequence will be concatenated and passed through an MLP which outputs a scalar value indicating how well the generated multi-view images align with human preferences.

### 4.2 MVReward Training

We train MVReward using a cross-entropy loss function, similar to those used in prior work (Stiennon et al. 2020) for RM training. For each comparison pair sampled from the human annotation dataset, associated with the same image prompt $I$ and two sequences of generated views $s_w = (v_w^{(0)}, v_w^{(1)}, \ldots, v_w^{(n)})$, $s_l = (v_l^{(0)}, v_l^{(1)}, \ldots, v_l^{(n)})$, if the sequence $s_w = (v_w^{(0)}, v_w^{(1)}, \ldots, v_w^{(n)})$ is better aligned with $I$, the loss function will be formulated as:

$$\mathcal{L}_\gamma = -\mathbf{E}_{(I,s_w,s_l)\sim\mathcal{H}}[\log(\sigma(r_\gamma(I, s_w) - r_\gamma(I, s_l)] \quad (1)$$

where $\gamma$ denotes the learnable parameters of MVReward, and $\mathcal{H}$ represents the training human annotation dataset.

We further enhance the model's domain-aware evaluation by distinguishing between features from different modalities (e.g., RGB and normal) and providing targeted feedback. Positional encoding is used to correlate the generated view modality with its position in the input sequence. We also introduce a set of modality-reversed negative samples, where the RGB and normal modalities are swapped. This contrastive approach enables the model to better understand the impact of positional changes across modalities, improving its modality-aware capabilities.

## 5 MVP: Multi-View Preference Learning

Building on MVReward's ability to encode human preferences for multi-view images, we further propose a plug-and-play tuning strategy to enhance the performance of multi-view diffusion models in aligning with human preferences.

### 5.1 Preliminaries

**Multi-view diffusion training.** Most multi-view diffusion models are fine-tuned from 2D diffusion models (DM) which lack awareness of image-view correlations. With the advent of large-scale 3D datasets (Deitke et al. 2023; Brazil et al. 2023), 3D priors have become enough for DMs to perceive the underlying correlations across different views. Existing multi-view DMs typically leverage the multi-view attention mechanisms to enhance consistency among views. Many methods like Wonder3D (Long et al. 2024), push a step forward by adding cross-domain attention to extend the generation to the normal domain, enabling the model to generate both color images and normal maps simultaneously. During the training process, the same noise is added to different views and domains (in Wonder3D, it's 12 images) of the same object. The DM UNet then predicts the noise across views, maintaining the L2 loss between predicted and added noise used in 2D diffusion.

**Reward feedback learning (ReFL).** ReFL, as proposed by ImageReward (Xu et al. 2024b), fine-tunes a text-to-image latent diffusion model (LDM) by predicting $x_t \rightarrow x'_0$ during the denoising steps. Empirical experiments show that the reward scores for generation $x'_0$ between 75%-99% of denoising steps are reliable for feedback learning. The mid-step t is randomly selected instead of using the last step to avoid an unstable fine-tuning process. After re-weighting the reward loss with the pre-training loss, ReFL demonstrates its effectiveness through impressive visual results.

Algorithm 1: Multi-View Preference Learning (MVP) for Multi-View DMs

**Pre-training Dataset:** Image-MV dataset $\mathcal{D} = \{(\text{img}_1, \text{mv}_1), ..., (\text{img}_n, \text{mv}_n)\}$, **where** $\text{mv}_i$= 6×[RGB+normal]$_i$
**Input:** Multi-view DM UNet with pre-trained parameters $w_0$, reward model $r$, multi-view DM pre-training loss function $\phi$, reward-to-loss map function $\psi$, re-weighting scale $\lambda$
**Initialization:** The noise scheduler $N_s$ and VAE of multi-view DM and time steps $\mathcal{T}$

**for** $(\text{img}_i, \text{mv}_i) \in \mathcal{D}$ **do**
  $\varepsilon \sim \mathcal{N}(0, I)$ // sample noise to be added to mv
  latent $\leftarrow \text{mv}_i + \varepsilon$ [with $N_s$, $\mathcal{T}$]
  embeds $\leftarrow$ pre-training process
  $\varepsilon' \leftarrow \text{UNet}_{w_i}(\text{latent}, \text{embeds}, \mathcal{T})$ // predict the noise
  $\mathcal{L}_{pt} \leftarrow \phi\,(\varepsilon, \varepsilon')$ // pre-training loss
  $\text{mv}'_i \leftarrow$ latent - $\varepsilon'$ [with $N_s$, $\mathcal{T}$, VAE]
  $\mathcal{L}_{rm} \leftarrow \psi\,(r\,(\text{img}_i, \text{mv}'_i))$ // reward loss
  $w_{i+1} \leftarrow w_i$ // update UNet [with $\lambda\mathcal{L}_{pt}$+$\mathcal{L}_{rm}$]
**end for**

### 5.2 Multi-View Preference Learning

Inspired by ReFL, we present Multi-View Preference Learning (MVP) to fine-tune multi-view diffusion models using MVReward. Following the standard multi-view diffusion pre-training, we add the same random noise to all ground-truth multi-view images and predict it with the pre-trained UNet, retaining the L2 loss between predicted and added noise. We then estimate the original latent through one-step sampling and decode them into images using the pre-trained VAE. The images are then fed into MVReward with their image prompt to obtain a reward score, which is converted into a reward loss through a reward-to-loss function map $\psi$. This reward loss is combined with the pre-training loss during backpropagation.

In practice, we observe significant generation deviations from the image prompt when simply adding reward loss to the pre-trained loss. Thus similar to ReFL, we introduce a large re-weighting scale $\lambda$ to the pre-training loss, ensuring strong pre-training constraints. The complete process is summarized in Algorithm 1, with the final loss function:

$$\begin{aligned} \mathcal{L} &= \lambda\mathcal{L}_{pt} + \mathcal{L}_{rm} \\ \mathcal{L}_{rm} &= \mathbf{E}_{\text{img}_i\sim\mathcal{D}}(\phi(r(\text{img}_i, g_\theta(\text{img}_i)))) \\ \mathcal{L}_{pt} &= \mathbf{E}_{(\text{img}_i,\text{mv}_i)\sim\mathcal{D}}(E_{\epsilon(\text{mv}_i),\text{img}_i,\epsilon\sim N(0,1),t} \\ &\qquad [\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(\text{img}_i))\|_2^2]) \end{aligned} \quad (2)$$

where $\theta$ is the multi-view DM parameters, and $g_\theta(\text{img}_i)$ denotes the predicted $\text{mv}_i$ from the multi-view DM with $\theta$ and image prompt. $\mathcal{L}_{pt}$ depends on the specific multi-view diffusion model to be fine-tuned. In our case it's the loss function derived from (Rombach et al. 2022).

## 6 Experiments

In this section, we present the quantitative and qualitative comparisons of MVReward and MVP against other existing methods to demonstrate their superiority. Limited by space, additional experimental results can be found in Appendix D.

| Metrics & Model | Human Eval. | | MVReward | | GSO Dataset | | CLIP | | BLIP | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Rank | Favor (%) | Rank | Reward | Rank | LPIPS ↓ | Rank | Score | Rank | Score |
| Envision3D | 7 | 0.06 | 7 | 0.205 | 3 | 0.130 | 6 | 0.774 | 4 | -0.253 |
| SyncDreamer | 6 | 0.10 | 6 | 0.690 | 7 | 0.146 | 7 | 0.759 | 2 | 0.057 |
| Wonder3D | 5 | 9.33 | 5 | 1.016 | 5 | 0.141 | 5 | 0.785 | 7 | -0.262 |
| - Wonder3D$_{pt}$ | - | - | - | 1.018 | - | 0.141 | - | 0.790 | - | -0.264 |
| - Wonder3D$_{mvp}$ (ours) | 4 | 13.7 | 4 | 1.389 | 6 | 0.142 | 4 | 0.796 | 5 | -0.258 |
| Zero123++ | 3 | 21.0 | 3 | 1.475 | 4 | 0.133 | 1 | **0.822** | 1 | **0.886** |
| Era3D | 2 | 25.3 | 2 | 1.506 | 2 | 0.126 | 3 | 0.801 | 6 | -0.259 |
| - Era3D$_{pt}$ | - | - | - | 1.503 | - | 0.127 | - | 0.802 | - | -0.259 |
| - Era3D$_{mvp}$ (ours) | 1 | **29.0** | 1 | **1.676** | 1 | **0.124** | 2 | 0.810 | 3 | -0.248 |
| Spearman to Human Eval. | - | | **1.00** | | 0.61 | | 0.86 | | 0.04 | |

Table 1: Quantitative comparisons of MVReward and MVP. Our MVReward outperforms all other metrics in aligning with human preferences, and our MVP consistently improves the performance of fine-tuned methods. The **bold** is the best results under certain metrics and the gray represents our results.

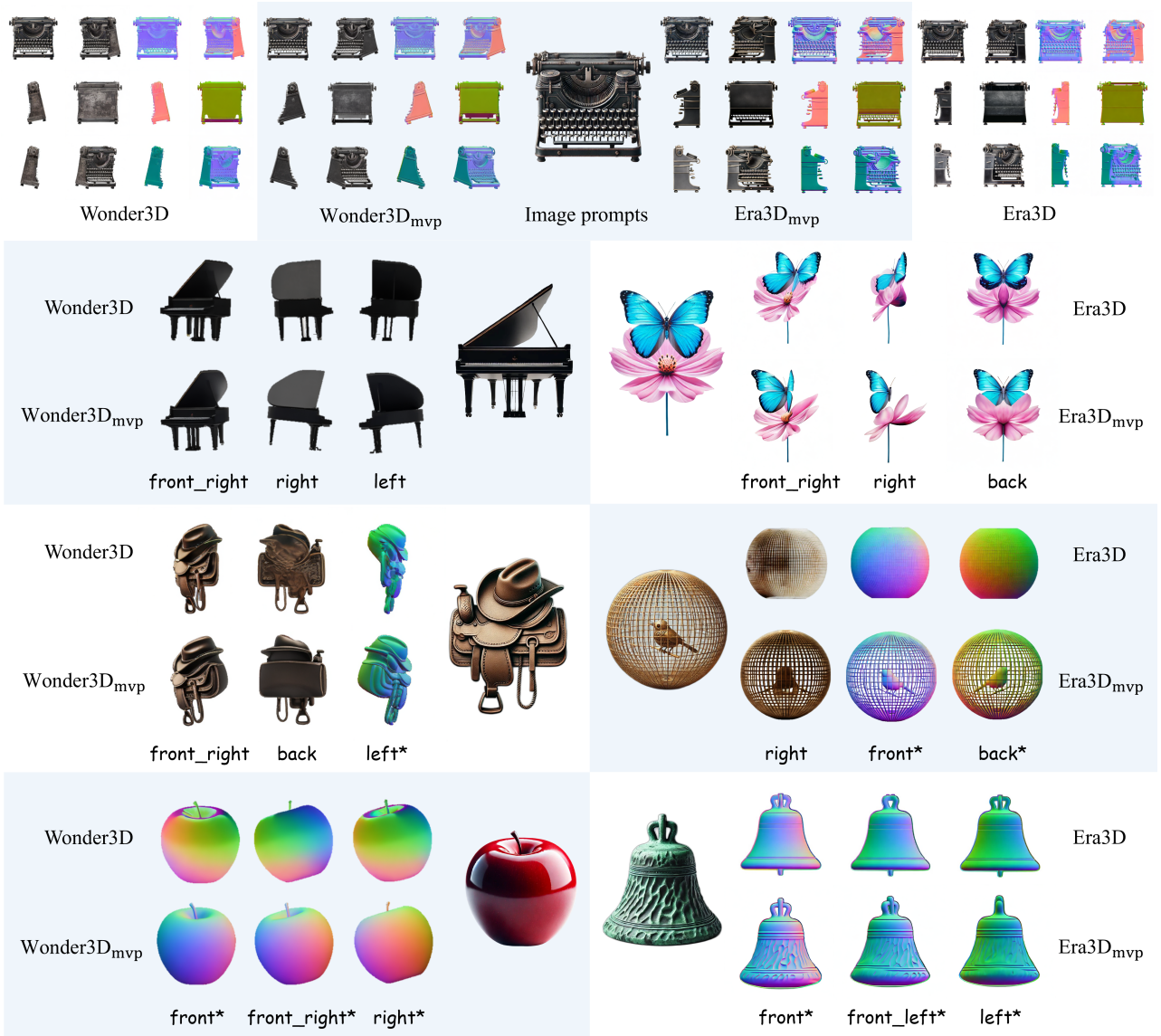


Figure 6: Qualitative comparisons between the original and the MVP fine-tuned multi-view diffusion models.

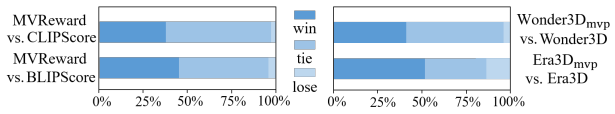


Figure 7: Win rates of MVReward and MVP. Our methods remain unbeaten versus their opponents most of the time.

### 6.1 MVReward: Predicting Human Preferences and Evaluating Multi-View Images

**Dataset & Training settings.** MVReward is trained on the human annotation dataset constructed in Section 3, with 1200 multi-view images asset-lists and 16k comparison pairs in total. The training, validation and test datasets are split according to an 8:1:1 ratio. The multi-view encoder is initialized using the pre-trained BLIP/VIT-B checkpoint. To prevent overfitting and ensure training stability, we fixed 50% of its parameters. Optimal performance is achieved with a batch size of 96 in total, an initial learning rate of 4e-5 using cosine annealing, on 4 NVIDIA Quadro RTX 8000.

**Quantitative results.** We conduct a user study to evaluate MVReward's ability in predicting human preferences. We collect multi-view assets from 7 methods on our standardized image prompt set and ask 6 new annotators to choose their favorite. For methods incapable of generating normal maps, we use GeoWizard (Fu et al. 2024) for prediction. Table 1 shows that our MVReward outperforms the commonly used GSO Dataset (LPIPS), CLIP Score, and BLIP Score in aligning with human preferences.

**Win rates.** We also calculate the win rates of MVReward versus other metrics in Figure 7(a). For each pairwise comparison, a metric wins if its preference aligns with human's while the opponent does not. A tie is recorded if two metrics have the same preference, which may occur frequently. Despite this, our MVReward still maintains a high win rate, demonstrating its strong alignment with human preferences.

**Ablation study.** We perform ablation studies on the encoder backbone, multi-view self-attention, and negative samples to assess their effects on MVReward. Table 2 indicates that substituting either results in reduced accuracy.

### 6.2 MVP: Aligning Multi-View Diffusion Models with Human Preferences

**Baseline & Training settings.** We select Wonder3D as our fine-tuning baseline since it is the only one of the four multi-view diffusion methods we used that has a publicly accessible training code. However, we also provide an unofficial fine-tuned version of Era3D to verify that MVP is a plug-and-play strategy. The model parameters are fixed except for the designated trainable modules within the UNet. Both models are fine-tuned in half-precision on 8 NVIDIA Quadro RTX 8000, with a batch size of 128 in total and a learning rate of 5e-6 with warm-up.

**Quantitative results.** For a fair comparison, we continue training the two pre-trained models with only pre-trained loss for the same iterations as fine-tuning process, namely Wonder3D$_{pt}$ and Era3D$_{pt}$ in Table 1. Since their results are nearly identical to the original versions, we exclude them from the user study and rankings, reporting only their metric scores. Our fine-tuned models Wonder3D$_{mvp}$ and Era3D$_{mvp}$ outperform both the original and pt-only versions across all the metrics, demonstrating the effectiveness of our MVP.

| | w/o BLIP | w/o MV SA | w/o Neg. | Full (ours) |
|---|---|---|---|---|
| Acc. | 81.7 | 81.5 | 74.8 | **83.1** |

Table 2: Ablation study of MVReward on backbone, multi-view self-attention (MV SA) and negative samples (Neg.), where w/o BLIP means changing the backbone to CLIP and Acc. denotes the evaluation accuracy of the reward model.

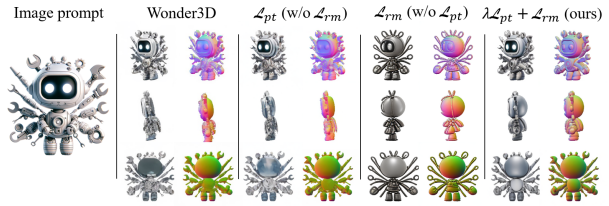


Figure 8: Further ablation experiment on the pre-trained loss and reward loss.

**Qualitative results.** Figure 6 presents several representative visual examples, highlighting the geometry and texture improvements brought by MVP (* means the normal map).

**Win rates.** Figure 7(b) illustrates the win rates of our fine-tuned models versus their original version during user study. Though tie occurs at times for simple objects, MVP consistently improves the generated multi-view images quality.

**Ablation study.** The comparison between the fine-tuned and pt-only version is an ablation study on the reward loss and we further develop it in Figure 8, illustrating that both the pre-trained loss and reward loss are essential in MVP.

## 7 Conclusion

In this paper, we address the challenges in aligning and evaluating multi-view diffusion models with human preferences. We construct a human annotation dataset and a standardized image prompt set through a comprehensive pipeline. We then train MVReward, an effective reward model that can serve as a reliable image-to-3D metric. We also introduce MVP as a plug-and-play tuning strategy for multi-view diffusion models to enhance their performance in human preference alignment. Experimental results support our contributions in providing a fair evaluation method and an efficient improving approach for multi-view diffusion models.

**Limitations.** Although fruitful, we admit that our reward model requires more training data for better evaluation. Additionally, focusing on multi-view images instead of direct 3D representations (i.e. mesh) limits the scope of our approach. We will endeavor to tackle these issues in the future and look forward to more potential insights.

## Acknowledgments

This research was funded through the National Key Research and Development Program of China (Project No. 2022YFB36066).

## References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Brazil, G.; Kumar, A.; Straub, J.; Ravi, N.; Johnson, J.; and Gkioxari, G. 2023. Omni3d: A large benchmark and model for 3d object detection in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13154–13164.

Chen, R.; Chen, Y.; Jiao, N.; and Jia, K. 2023. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 22246–22256.

Chen, Z.; Wang, Y.; Wang, F.; Wang, Z.; and Liu, H. 2024. V3d: Video diffusion models are effective 3d generators. *arXiv preprint arXiv:2403.06738*.

Choi, J. S.; Lee, K.; Lee, D.; Shin, J.; and Lee, K. 2024. HFDream: Improving 3D Generation via Human-Assisted Multi-view Text-to-Image Models.

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Deitke, M.; Schwenk, D.; Salvador, J.; Weihs, L.; Michel, O.; VanderBilt, E.; Schmidt, L.; Ehsani, K.; Kembhavi, A.; and Farhadi, A. 2023. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13142–13153.

Downs, L.; Francis, A.; Koenig, N.; Kinman, B.; Hickman, R.; Reymann, K.; McHugh, T. B.; and Vanhoucke, V. 2022. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, 2553–2560. IEEE.

Fu, X.; Yin, W.; Hu, M.; Wang, K.; Ma, Y.; Tan, P.; Shen, S.; Lin, D.; and Long, X. 2024. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. *arXiv preprint arXiv:2403.12013*.

He, Y.; Bai, Y.; Lin, M.; Zhao, W.; Hu, Y.; Sheng, J.; Yi, R.; Li, J.; and Liu, Y.-J. 2023. T $^3$ Bench: Benchmarking Current Progress in Text-to-3D Generation. *arXiv preprint arXiv:2310.02977*.

Hong, Y.; Zhang, K.; Gu, J.; Bi, S.; Zhou, Y.; Liu, D.; Liu, F.; Sunkavalli, K.; Bui, T.; and Tan, H. 2023. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*.

Ibarz, B.; Leike, J.; Pohlen, T.; Irving, G.; Legg, S.; and Amodei, D. 2018. Reward learning from human preferences and demonstrations in atari. *Advances in neural information processing systems*, 31.

Kirstain, Y.; Polyak, A.; Singer, U.; Matiana, S.; Penna, J.; and Levy, O. 2023. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36: 36652–36663.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.

Li, J.; Tan, H.; Zhang, K.; Xu, Z.; Luan, F.; Xu, Y.; Hong, Y.; Sunkavalli, K.; Shakhnarovich, G.; and Bi, S. 2023. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. *arXiv preprint arXiv:2311.06214*.

Li, P.; Liu, Y.; Long, X.; Zhang, F.; Lin, C.; Li, M.; Qi, X.; Zhang, S.; Luo, W.; Tan, P.; et al. 2024. Era3D: High-Resolution Multiview Diffusion using Efficient Row-wise Attention. *arXiv preprint arXiv:2405.11616*.

Lin, C.-H.; Gao, J.; Tang, L.; Takikawa, T.; Zeng, X.; Huang, X.; Kreis, K.; Fidler, S.; Liu, M.-Y.; and Lin, T.-Y. 2023. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 300–309.

Liu, R.; Wu, R.; Van Hoorick, B.; Tokmakov, P.; Zakharov, S.; and Vondrick, C. 2023a. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9298–9309.

Liu, Y.; Lin, C.; Zeng, Z.; Long, X.; Liu, L.; Komura, T.; and Wang, W. 2023b. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*.

Long, X.; Guo, Y.-C.; Lin, C.; Liu, Y.; Dou, Z.; Liu, L.; Ma, Y.; Zhang, S.-H.; Habermann, M.; Theobalt, C.; et al. 2024. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9970–9980.

Otani, M.; Togashi, R.; Sawai, Y.; Ishigami, R.; Nakashima, Y.; Rahtu, E.; Heikkilä, J.; and Satoh, S. 2023. Toward verifiable and reproducible human evaluation for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14277–14286.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Pang, Y.; Jia, T.; Shi, Y.; Tang, Z.; Zhang, J.; Cheng, X.; Zhou, X.; Tay, F. E.; and Yuan, L. 2024. Envision3D: One Image to 3D with Anchor Views Interpolation. *arXiv preprint arXiv:2403.08902*.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, 8821–8831. Pmlr.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Shi, R.; Chen, H.; Zhang, Z.; Liu, M.; Xu, C.; Wei, X.; Chen, L.; Zeng, C.; and Su, H. 2023. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*.

Stiennon, N.; Ouyang, L.; Wu, J.; Ziegler, D.; Lowe, R.; Voss, C.; Radford, A.; Amodei, D.; and Christiano, P. F. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33: 3008–3021.

Tang, J.; Chen, Z.; Chen, X.; Wang, T.; Zeng, G.; and Liu, Z. 2024. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. *arXiv preprint arXiv:2402.05054*.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Voleti, V.; Yao, C.-H.; Boss, M.; Letts, A.; Pankratz, D.; Tochilkin, D.; Laforte, C.; Rombach, R.; and Jampani, V. 2024. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. *arXiv preprint arXiv:2403.12008*.

Wang, H.; Du, X.; Li, J.; Yeh, R. A.; and Shakhnarovich, G. 2023. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12619–12629.

Wang, Z.; Lu, C.; Wang, Y.; Bao, F.; Li, C.; Su, H.; and Zhu, J. 2024a. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36.

Wang, Z.; Wang, Y.; Chen, Y.; Xiang, C.; Chen, S.; Yu, D.; Li, C.; Su, H.; and Zhu, J. 2024b. Crm: Single image to 3d textured mesh with convolutional reconstruction model. *arXiv preprint arXiv:2403.05034*.

Wu, T.; Yang, G.; Li, Z.; Zhang, K.; Liu, Z.; Guibas, L.; Lin, D.; and Wetzstein, G. 2024. Gpt-4v (ision) is a human-aligned evaluator for text-to-3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22227–22238.

Wu, X.; Sun, K.; Zhu, F.; Zhao, R.; and Li, H. 2023. Human preference score: Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2096–2105.

Xu, J.; Cheng, W.; Gao, Y.; Wang, X.; Gao, S.; and Shan, Y. 2024a. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*.

Xu, J.; Liu, X.; Wu, Y.; Tong, Y.; Li, Q.; Ding, M.; Tang, J.; and Dong, Y. 2024b. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36.

Xu, Y.; Shi, Z.; Yifan, W.; Chen, H.; Yang, C.; Peng, S.; Shen, Y.; and Wetzstein, G. 2024c. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. *arXiv preprint arXiv:2403.14621*.

Ye, J.; Liu, F.; Li, Q.; Wang, Z.; Wang, Y.; Wang, X.; Duan, Y.; and Zhu, J. 2024. Dreamreward: Text-to-3d generation with human preference. *arXiv preprint arXiv:2403.14613*.